DOCUMENT RESUME

ED 416 453 CS 013 058

AUTHOR Maxwell, Martha
TITLE Evaluating Individualized Reading Programs: A Bayesian Model.
PUB DATE 1998-04-03
NOTE 12p.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Data Collection; Decision Making; Higher Education; Probability; Program Effectiveness; *Reading Programs; *Research Methodology; Standardized Tests; Statistical Inference; Study Skills
IDENTIFIERS Variables

ABSTRACT

Simple Bayesian approaches can be applied to answer specific questions in evaluating an individualized reading program. A small reading and study skills program located in the counseling center of a major research university collected and compiled data on student characteristics such as class, number of sessions attended, grade point average, and other demographic characteristics. However, there is no valid way to draw conclusions across such variables. A more meaningful way to present data of this type is to construct a probability tree. Using parametric statistics like means, and standard deviations, correlations require that certain assumptions be met (interval measurement, normal distributions, homogeneity of variance, some variance to begin with, etc.). Standardized reading tests are not adequate criteria of either reading program effectiveness nor do they reflect the reading demands of college courses realistically. Attendance can be a useful criteria for measuring a program's effectiveness. Bayesian technique as applied to decision-making implies that evaluation is a continuous process, and that evaluation is not necessarily concerned with generating new knowledge nor finding ultimate truths which may be the goals of the researcher. Such techniques, used appropriately, can eliminate the expense and effort of gathering of masses of data over a long period of time to make decisions. Arranging demographic and outcome data in Bayesian probability trees makes data easier to understand and interpret. (Contains 11 references and 3 tables of data.) (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 416 453

Evaluating Individualized Reading Programs: A Bayesian Model.

Martha Maxwell (3/ 4,/98)
Box 2857
Kensington, MD 20891

College adminstrators in their attempts to preserve present student service programs including those that provide academic support are facing new fiscal challenges and restrictions. In the future, most public institutions are expected to serve more students with less money as they face increases in student enrollment coupled with fewer dollars from federal and state funding. Traditional private colleges also face budgetary problems as new commercial universities and distance education facilities compete with them for students.

As a result, administrators are demanding more accountability from academic support programs and requiring them to provide more quantitative data showing that their outcomes are worth the investment. Whether you call it bean counting or number crunching, administrators today want quantitative evidence that programs are working well and doing what they purport to do.

Although researchers have been complaining for decades about the paucity of program evaluation efforts in academic support programs (Roueche, 1968, Donovan, 1975, Black & others, 1991, Boylan and others, 1994), recent studies suggest that only about 20 percent of the programs actively engage in systematic evaluation that describe how well it does what it does and what it does well (Boylan,1997).

Boylan (1997) contends that although we now have clear evidence that well-designed and properly-implemented developmental programs can improve student retention, grades and graduation rates, today we need more program evaluation to expand our knowledge base about the specific activities that contribute to that success and who is most likely to benefit from these activities.(Boylan, 1997). In other words, we need to know which particular programs or techniques work well with which students.

Boylan (1997) emphasizes that rather than using experimental research techniques, program research or evaluation involves simple descriptive techniques to look at what works well and to determine who benefits. As he points out , "in most cases , the use of percentages,- averages, pie charts or frequency distributions is sufficient to analyze and present the information resulting from program research, and that by analyzing the percentage of various demographic groups who used the services, one might later track the performance of these students to see how long they were retained or the grades they received in subsequent courses. " (p. 27)

CS013058

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Maxwell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

But tables, pie charts and frequency distributions also have disadvantages and limitations for they don't reveal the interactions between variables and can lead to incorrect assumptions and inappropriate decisions. One way of systematizing these data that will greatly clarify the questions we need to answer is to use simple Bayesian analysis (Maxwell 1970-71).

An illustration of how Bayesian approaches can be applied to answer specific questions in evaluating an individualized reading program is a study by Maxwell, 1970-71. The study we describe involves a small reading and study skills program located in the counseling center of a major research university. Housed in an old WWII temporary officers' barracks, the program, through its advertising on campus, attracted a wide variety of clients. Students were not charged for the program, attendance was voluntary and no credit was given. Students entering the program were interviewed and depending on their needs and interests were given an individualized program.

Like other student services, the program collected and compiled data on student characteristics such as class, number of sessions attended, grade point average, and other demographic characteristics. Table 1 shows the data compiled on three characteristics of reading program students.

INSERT TABLE 1 ABOUT HERE

The distributions shown in Table 1 are typical of those used to describe student clientele in terms of descriptive statistics (means, median, mode) and interpretations are often made across the variables since the same individuals are involved in each distribution.

However, each variable listed in Table 1 must be viewed as a mutually exclusive event in that a given individual can only appear in one cell at a time-- that is, a student cannot be both a sophomore and a graduate student simultaneously. Since properties of mutually exclusive events are additive, the percentages of cases in each variable total 100%. If we are concerned with about predicting how many of the next 200 students who request tutoring are likely to be graduate students, we can use the above percentages as probabilities and predict that 30 percent or 60 of the next two hundred clients seen will probably be graduate students. Because we cannot be completely certain about what will happen in the future - that is, if we are predicting under conditions of uncertainty, we can use our empirical sample frequencies as the basis for making predictions in probabilistic terms. Similarly, we could estimate the numbers of students likely to fall into the other sub-groups - e.g., language background, and GPA.

However, there is no valid way to draw conclusions across variables arranged as they are in Table 1. In fact, such distributions encourage the reader to resort to his/her own biases and prejudices in forming conclusions. A college administrator

TABLE 1
Characteristics of Students Using the Reading and Study Skills Service

Class	N	%	Language Background	N	%	Grade Point Average	N	%
Graduate	30	16						
Senior	31	17	English speaking	109	59	3.5 +	30	16
Junior	47	25				3.0 — 3.4	44	23
Sophomore	29	15	Non-English speaking	77	41	2.9 or lower	55	30
Freshman	49	27				None (new students)	57	31
	186	100		186	100		186	100

looking at Table 1 might express surprise that there were so many graduate students seeking help in reading and conclude that :"Of course, they all must be foreign students who are having problems with English and are making poor grades in their courses."

These kinds of interpretations are analogous to trying to describe a pile of objects consisting of 6 red apples, 3 green apples, 5 walnuts, 4 brown rocks, and 2 diamonds by selecting three characteristics - name of object, color and hardness and concluding that the typical object in the pile is a brown, hard (unbreakable) apple.

A more meaningful way to present data of this type is to construct a *probability tree* so that the relevant characteristics of subgroups across variables can be readily observed as follows:

<INSERT TABLE 2 ABOUT HERE.>

Note: Non-English Background refers to foreign students and minorities (Hispanic Americans and Chinese whose English is their second language.)

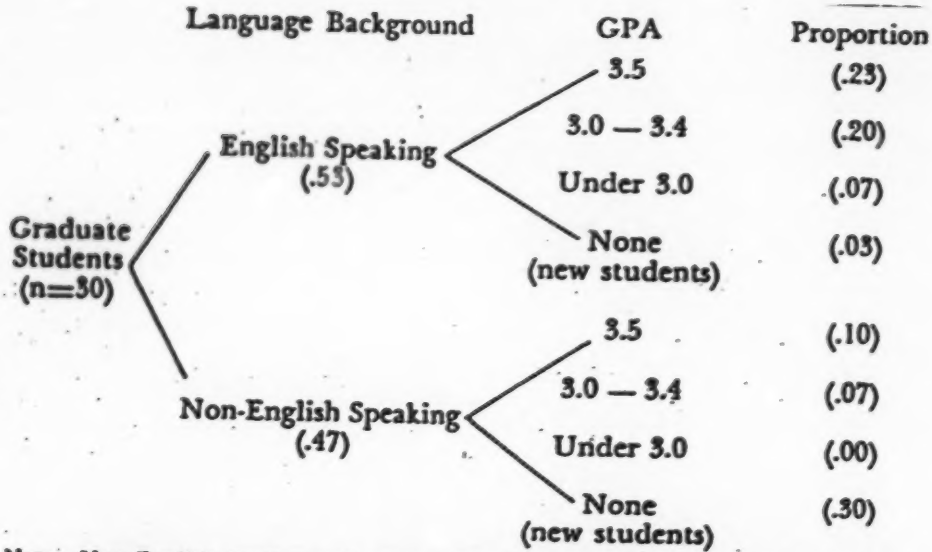
Our college administrator, looking at Table 2 would not draw the conclusion cited above , for this breakdown shows that none of the graduate students had low GPA's. However, 30% of the graduate group were from non-English speaking backgrounds but as yet had no GPA's, suggesting that they were new to the institution and *may* have been anxious about their ability to adapt to graduate work. This hypothesis would need further testing.

Were we to use traditional research methodology and statistics to evaluate the effectiveness of our reading program, we might administer a reading pre-test to the total group, give them treatment (i.e., an individualized reading program) and then test them again at the end of the program --either testing the entire group or selecting a random sample. In either case, we would undoubtedly lose cases between the program's beginning and end and would face the problem of how to handle the data statistically - i.e., finding statistics suitable for unequal numbers can be a bothersome task. We might endeavor to find a control group with whom we could compare our reading students and either pull a random sample of students who were not in the program or try to match each student in our program with someone who was not in the program. Inevitably, we would find the experimental and control groups differed on some important variable (like SAT scores) and resort to covariance techniques to attempt to control for individual differences.

Even then our study would be criticized for not controlling on motivation under the assumption that those who voluntarily seek help in reading are different from those who do not. We could try to avoid this criticism by withholding the reading program from a group of students and making them wait until the treatment program was over and then trying to cajole the non--treatment group into taking

BEST COPY AVAILABLE

TABLE 2
Bayesian Tree Showing Characteristics of Graduate Student Clients



Note: Non-English Background refers to foreign students and minorities (Mexican American and Chinese whose English is their second language).

the post-tests at the same time as the experimental group. In addition to the many problems involved in trying to set up this study in a practical situations, there are a number of assumptions underlying traditional experimental design and statistics which may not be relevant for evaluating an on-going individualized reading program.

Using parametric statistics like means, standard deviations, correlations, etc. requires that certain assumptions be met (i.e., interval measurement, normal distributions, homogeneity of variance, some variance to begin with, etc.) Starting a study with a large initial N is almost always a "must" affecting even the use of non-parametric measures such as chi-square. This might lead us to classify students into illogical groups (like combining rotten apples and diamonds).

To illustrate another aspect of the problems involved in applying parametric statistics to our example, consider the descriptions of the first five students pulled in a random sample:

> **1. A law student from Malaysia whose reading rate on easy material was 190 words per minute and who wanted help in reading a text in tax law.**
>
> **2. A junior literature major with a 4.0 average whose reading rate was 450 words per minute who wanted to double her speed so she could devote more time to the novel she was writing.**
>
> **3. A Chinese-American sophomore with A's in physics and math who was failing sociology. His reading rate on easy material was 200 words per minute with 50% comprehension.**
>
> **4. A freshman (honor student in high school) who was making D's in most of his courses because he was unable to compete more than a third of the reading his professors assigned.**
>
> **5. A dance major whose reading rate on literature was 350 words per minute with 90% comprehension, but who read only 150 words per minute on technical material with 40% comprehension and was having difficulty reading her physiology textbook.**

Even if we could be sure that our pre-test adequately reflected these students' reading abilities, it is hard to envision a reading program that would meet each of their needs. Nor could we reasonably expect to find statistically significant results by administering a standardized post-test. Presently, standardized reading tests are not adequate criteria of either reading program effectiveness nor do they reflect the reading demands of college courses realistically, therefore, we must search for other criteria.

BEST COPY AVAILABLE

Attendance as a Criteria of Program Effectiveness

Attendance can be a useful criteria for measuring a program's effectiveness as numerous early studies have shown (Maxwell, 1970-71). The various rationales for using persistence as the criterion include the fact that college students have many demands on their time so if they continues working on a voluntary program, they probably feel that it is meeting their needs, that they are making progress toward their goals, that they see the reading service as relevant to their college courses and they are motivated to improve their skills.

In an earlier study, Maxwell (1969) found that students who remained in a voluntary individualized reading program for six weeks (12 hours of practice) showed greater gains on a post-test than those who spent less time. She observed that it seemed to take about six weeks for the students to internalize the new skills they were learning. Also she observed that many students neither want nor need an intensive reading program -- that is, some need help in coping with a specific assignment, or want to discuss their anxieties about an exam or assignment, etc. Thus, the reading specialist's role in an individualized program is to diagnose the student's difficulty, determine the appropriate services needed, plan individual programs should s/he enter the individualized program, help the student set goals, and monitor his/her progress.

<INSERT TABLE 3 ABOUT HERE>

The probability tree in Table 3 shows how class, language background, and type of service relate to attendance in the program. One thing stands out clearly in Table 3- that is, freshmen and graduate students from non-English speaking backgrounds are more likely to remain in the program for six weeks than juniors and seniors for whom English is a second language. This raises a number of questions: Are the drop-outs being adequately diagnosed by the reading specialist? Is the service appropriate for them? Are their needs different- for example, are freshmen and beginning graduate students more anxious about starting in a new institution? How well is the present program meeting the needs of those who drop out? We might need conferences or resort to other strategies to answer these questions.

After considering Table 3, we might decide to explore the breakdown further and alter the program to see if it makes a difference. For instance, let us assume that we find that 70% of those assigned individual programs requested help in improving their reading and comprehension skills. However, we also find that these same students were taking courses requiring extensive term papers and they needed to become more skilled in pulling out generalizations and facts from their reading and organizing them around a thesis statement. With this information, the director could consider ways of altering the program to better meet these needs-- perhaps by offering a short mini-course on techniques for finding main ideas, developing a thesis statement, and organizing ideas. If this mini-course were offered, the number

BEST COPY AVAILABLE

TABLE 3

Bayesian Tree Showing Type and Amount of Service Used and Characteristics of Students Seeking Help During the First 6 Weeks of the Term

Class	Language Background	Type of Service	Proportion Remaining in Lab Group For 6 Weeks
Graduate (.16)	Non-English (.53)	Interviews only (.50)	
		Interview + lab (.50)	.75
	English (.47)	Interviews only (.64)	
		Interview + lab (.36)	.40
Senior (.17)	Non-English (.33)	Interviews only (.40)	
		Interview + lab (.60)	.10
	English (.66)	Interviews only (.48)	
		Interview + lab (.52)	.27
Junior (.25)	Non-English (.32)	Interviews only (.50)	
		Interview + lab (.50)	.11
	English (.68)	Interviews only (.52)	
		Interview + lab (.48)	.36
Sophomore (.15)	Non-English (.45)	Interviews only (.77)	
		Interview + lab (.23)	.33
	English (.55)	Interviews only (.63)	
		Interview + lab (.37)	.50
Freshman (.27)	Non-English (.45)	Interviews only (.59)	
		Interview + lab (.41)	.78
	English (.55)	Interviews only (.63)	.60
		Interview + lab (.37)	

BEST COPY AVAILABLE

of students starting individualized programs might be smaller , but we could compare the proportion staying for six weeks with the percentages in Figure 3 to determine whether offering students the mini-course increased the percentage who remained for six weeks in the individualized program. It is true, that we'd expect the percentage of those in individual programs to be smaller but the percentage of those who continued for six weeks might increase. The attendance percentages from Table 3 could be used as probabilities.

If decisions to alter the program's routine are implemented, they are sometimes criticized by statisticians for being based on a small number of cases. Yet if we were managing a factory that manufactured personal computers which cost $700 apiece to produce, wee would not wait until we produced 100 of these only to find that 70 of the 100 were incorrectly assembled and had to be scrapped. If we waited, we would soon be bankrupt. Yet sometimes that high a percentage of students in our programs and/or classrooms drop out and/or we find that our teaching methods fail to yield statistically significant changes.

Bayesian technique as applied to decision-making imply that evaluation is a continuous process, and that evaluation is not necessarily concerned with generating new knowledge nor finding ultimate truths which may be the goals of the researcher. The decision maker needs methodology for collecting information, monitoring the program, and taking corrective action when needed. Although Bayesian statistics as presently used in psychometrics, computer science, and economics sometimes can be very complex and sophisticated (Novick & otthers, 1970; Novick, 1974, Hashway, 1998,) the decision trees we have illustrated are easy for a teacher to set up and interpret. Bayesian methods have been found to yield results equivalent to using traditional parametric statistics, on arrays of test scores to set probabilities. (Meyer, 1963.) Novick and Jackson (1970) for example, developed techniques for determining the different probabilities of a students succeeding college based on entrance test scores, probability of finishing the freshman year, expected grade-point-average depending on the type of institution s/he entered and his/her probability of completing college and overall GPA.

The Bayesian statistical base provides methods for making decisions when only a minimum amount of information is available and the numbers are small. It yields a powerful statistical method of evaluating new information and revising original estimates of the probability that events are in one state or another. If used, appropriately it can eliminate the expense and effort of gathering of masses of data over a long period of time in order to make decisions.

There are many other decisions in administering a reading program where Bayesian thinking might well be applied. Decisions involving staffing, training staff, planning for maximum use of equipment, materials, testing different teaching strategies might be based on Bayesian analysis. Directors of new services could use the experiences of others who conduct similar programs with similar populations to make subjective

probability estimates and have a basis for making wiser initial decisions. And probability trees could help identify the characteristics of sub-groups of students on whom more intensive research might be conducted.

Probability theory and the Bayesian model have rarely been applied to academic support or developmental education programs. These techniques, if appropriately used, could help us resolve some of our basic problems in evaluating many of our services including peer tutoring. Arranging demographic and outcome data in Bayesian trees will make our data easier to understand and interpret, and lead to insights that will enable us to improve our programs as well as justifyour existence.

REFERENCES

Black, M., Mansfield, W. & Farrin, E. (May 1991). *College Level Remedial Education in the Fall of 1989., Survey Report*. Washington, DC: U. S. Department of Education.

Boylan, H. R. (Fall 1997). The case for program research in developmental education. *The Learning Assistance Review, 2*(2), 20-34.

Boylan H .R., Bonham, B. S. & Bliss, L . B. (1994). Characteristics Components of Developmental Programs. *Research in Developmental Education, 11*(1).

Donovan, R. A. (1975). *National Project II: Alternatives to the revolving door.* New York: Bronx Community College.

Hashway, R. M. (1998). *Assessment and Evaluation of Developmental Learning: Qualitative Individual Assessment and Evaluation Models.* Westport, CN: Praeger (Bayesian Methods, pp 177-124.

Maxwell, M. J. (1971-2) Evaluating an individualized college reading service: Alternative criteria and statistical models to the test-retest paradigm. *Journal of Reading Behavior,4*(3), 1-8.

Maxwell, M. J. (1969). Predicting persistence in a voluntary college reading program. (Paper read at the International Reading Association Convention, 1969).

Meyer, D. L.(1961). Bayesian statistics. *Methodology of Educational Research, Review of Educational Research, 39,* 723-727.

Novick, M. R., Jackson, P. H. , Thayer, D. . & Cole N. S. (1971). *Applications of the Bayesian Model to the Prediction of Educational Performance. Monograph No. 42,* Circle Pines, IA: American College Testing.

Novick, M. R., (1974). ***Statistical Methods for Educational and Psychological Research.*** **New York, McGraw-Hill.**

Roueche, J. E. (1968). ***Salvage, redirection or custody.*** **Washington, D. C.: American Association of Junior Colleges.**